UDK: 811.163.42:81'374:81'32:81'24
Pregledni znanstveni rad
Prihvaćen za tisak: 10. prosinca 2013.

Gordana Hržica
Sveučilište u Zagrebu, Laboratorij za psiholingvistička istraživanja
ghrzica@erf.hr

Jelena Kuvač Kraljević
jkuvac@erf.hr

Jan Šnajder
Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
jan.snajder@fer.hr

Hrvatski čestotni rječnik dječjega jezika

*Jezični korpusi priznati su izvori jezičnih podataka. Međutim, dohvat tih podataka može biti složen i dugotrajan za krajnje korisnike. Hrvatski korpus dječjega jezika (*HKDJ *— Kovačević 2002) jedini je hrvatski korpus govornoga jezika. Sastoji se od prijepisa uzoraka spontanoga govornoga jezika troje djece. Djeca su uzorkovana u pravilnim vremenskim razmacima, od progovaranja do treće godine. Uzorci su transkribirani u programskom paketu* CLAN*, u skladu s pravilima* CHAT*-a. Dostupni su u Svjetskoj bazi dječjega jezika* CHILDES *(http://childes.psy.cmu.edu/data/Slavic/). Cilj je* HKDJ*-a pružiti podatke o leksičkom i gramatičkom razvoju u ranom jezičnom usvajanju. Kako bi se omogućio lakši i precizniji dohvat podataka dostupnih u* HKDJ*-u, pristupilo se izradi Hrvatskoga čestotnoga rječnika dječjega jezika (DjeČeR) čija je izrada još u tijeku. DjeČeR vjerodostojno odražava specifičnosti* HKDJ*-a (govorni korpus s razvojnom komponentom) te je sukladno strukturiran. U DjeČeRu je omogućen pregled natuknica triju potkorpusa* HKDJ*-a zasebno te unije i presjeka potkorpusa triju govornika. Pretražuje se prema čestotnosti, abecednom redu, vremenu pojave i vrsti riječi. U Dje-*

ČeRu su dostupni i podatci o morfosintaktičkom opisu natuknica[1] *koje se pojavljuju u* HKDJ*-u te točan popis njihovih obličnica.*

1. Uvod

Jezični je korpus prepoznat kao vrijedan i neiscrpan izvor jezičnih podataka još potkraj XIX. st., ali je njegova sustavna i objektivna uporaba započela sredinom šezdesetih godina prošloga stoljeća kada su Francis i Kučera (1967) oblikovali prvi računalni korpus. Karakteristike korpusa kao objektivnost, pristupačnost, provjerljivost, primjenljivost u dijakronijskim i sinkronijskim istraživanjima (Svartvik 1992), potvrdile su se u strukturalno i sadržajno različitim korpusima.

Iako vremenski gledano hrvatska korpusna lingvistika nije slijedila svjetsku, ipak postoji pozamašan opus hrvatskih korpusa počevši od 60-ih godina prošloga stoljeća pa do danas. Ti su korpusi, jednako kao i strani, imali različite ciljeve pa su se s obzirom na unaprijed postavljene ciljeve temeljili na različitoj jezičnoga građi. Tako je, primjerice, Furlanov korpus (1961) imao obrazovne ciljeva pa se shodno tome temeljio na tekstovima školskih udžbenika i učeničkim zapisima. Bujas je u stvaranju korpusa (1975.a, 1975.b) vidio dobar izvor za stvaranje leksikografske građe. Za razliku od engleskih korpusa, hrvatski su korpusi strukturom češći i dostupniji kao pisani nego kao korpusi govornoga jezika. Jedan od glavnih razloga leži u nešto jednostavnijemu oblikovanju pisanih korpusa. Stvaranje govornoga korpusa uz brojna pitanja i nedoumice koja se vežu uz pisane korpuse nosi i niz posebnih pitanja kao na primjer: koliko iskaza uključiti da bi prijepis bio vjerodostojan, treba li transkribirati po fonetskom ili ortografskome načelu, što činiti s govornim pojavama kao što su zastajkivanja, oklijevanja, ponavljanja i slično, govore li nam ona išta o jezičnoj obradi.

Svaki bi korpus trebao ponuditi tri vrste podataka (Tadić 2003): 1. evidenciju — ima li nečega u korpusu ili ne; 2. frekvenciju — čestotu neke jezične jedinice i 3. relaciju — odnos jedne jezične jedinice prema drugoj.

S obzirom na vrstu korpusa i vrstu informacije cilj je ovoga rada prikazati govorni korpus s posebnim osvrtom na čestotu leksičkih jedinica koje su u njemu sadržane. Trenutno u hrvatskome govornome području postoji samo jedan govorni korpus — *Hrvatski korpus dječjega jezika* (HKDJ — Ko-

[1] Naziv *natuknica* različito se tretira u dvije različite tradicije u jezikoslovlju, leksikografskoj i psiholingvističkoj (Jelaska, 2005). Unutar leksikografske tradicije označava osnovu unosa u rječnik, to jest početak rječničkoga članka (u tom se značenju u engleskome osim naziva *lemma* upotrebljava i naziv *headword*), a u psiholingvistici označava apstraktni osnovni oblik riječi (eng. *lemma*), to jest jedinicu umnoga rječnika (Crystal 2003).

vačević 2002). Govorni korpus jezika odraslih tek je u izradi, a kao pandan HKDJ-u nužan je ne samo kako bi se upotpunio okvir govornih korpusa, već kako bi se sinkronijska istraživanja dječjega jezika nadovezala na jezik odraslih. Leksičku čestotnu inačicu HKDJ-a ostvario je *Hrvatski čestotni rječnik dječjega jezika* (DjeČeR — Kuvač Kraljević, Hržica, Štefanec, u pripremi). No, da bi se razumjela struktura, a posebice način oblikovanja i nastanak DjeČeRa, prvo će se opisati HKDJ, a zatim pobliže prikazati leksička struktura DjeČeRa.

1.1. Hrvatski korpus dječjega jezika

Različite metode u prikupljanju jezičnih podataka pružaju različite obavijesti o jezičnome razvoju i obradi. Pitanje koju metodu upotrijebiti čest je metodološki kamen spoticanja istraživača u lingvističkim i psiholingvističkim istraživanjima. Nažalost, jednostavnoga odgovora nema, već ga određuje niz čimbenika kao što su uvjeti ispitivanja (ispitanici, oprema), duljina ispitivanja (longitudinalna nasuprot transverzalna ispitivanja), vrsta jezičnih podataka koja se želi prikupiti, vrijeme ispitivanja (npr. koje se razdoblje jezičnoga razvoja želi opisivati) te niz drugih. Različite kombinacije različitih istraživačkih varijabli osiguravaju različite podatke. Primjerice, longitudinalna ispitivanja, odnosno dugogodišnje praćenje većega broja ispitanika, pružaju sliku jezičnoga razvoja svakoga ispitanika posebno s mogućnošću promatranja individualnih razlika. Najčešći je nedostatak ove metode duljina praćenja, što znači da može proći mnogo godina do donošenja prvih zaključaka. Suprotno tome, transverzalna ispitivanja, odnosno ispitivanje većega broja sudionika u kraćem razdoblju, osiguravaju brži dolazak do podataka. Ovakva su ispitivanja češća u jezičnim istraživanjima jer se temelje na ispitivanju većega broja sudionika različite dobi, spola, kliničkih skupina i slično, na istom, prethodno metodološki razrađenome jezičnome materijalu. Međutim, nedostatci su im, primjerice, manje pouzdani opisi individualnih razlika, veća vjerojatnosti griješenja pri ispitivanju ili nemogućnost utjecanja na djetetovo pogađanje (Behrens 2008).

Kada se zadovolji niz uvjeta (dobro uvježbani ispitivači, longitudinalno praćenje, nekoliko sudionika koji su voljni sudjelovati u dugotrajnom ispitivanju, dobra audiovizualna oprema, pravilan vremenski interval praćenja, jasno određen sadržaj ispitivanja te način obrade slušno prikupljenih podataka), postavljeni su glavni preduvjeti za oblikovanje govornoga korpusa. Ti su preduvjeti bili osigurani početkom 90-ih godina prošloga stoljeća kada se pristupilo izradi Hrvatskoga korpusa dječjega jezika (HKDJ, Kovačević 2002) Korpus je nastao snimanjem spontanoga govora u svakodnevnim jezičnim situacijama troje jednojezične djece urednoga jezičnoga razvoja od njihove prve do treće godine života. Slijedeći zbivanja svjetske psiholingvis-

tike, nastanak je prvoga hrvatskoga dječjega govornoga korpusa potaknulo pokretanje Sustava za razmjenu podataka dječjega jezika poznatijega pod akronimom CHILDES (*Child Language Data Exchange System*), koji su 1981. pokrenuli Catherine Snow i Brian MacWhinney sa sjedištem u Pittsburghu (MacWhinney 2000). Svrha CHILDES-a bila je usustaviti prikupljene podatke dječjega jezika te tako usustavljenim korpusom stvoriti platformu za daljnja empirijska istraživanja, a u cilju prihvaćanja i opovrgavanja postojećih, pa čak i oblikovanja novih teorija. U tu svrhu osmišljen je niz pravila za transkripciju i kodiranje (CHAT — *Codes for the Human Analysis of Transcripts*) te osigurana računalna podrška (CLAN — *Computerized Language Analysis*) (MacWhinney i Snow, 1985).

Prvenstvena je svrha HKDJ-a pružiti sinkronijske i dijakronijske opise jezičnoga razvoja u hrvatskome do treće godine, primarno na morfološkoj, leksičkoj i sintaktičkoj razini. Nadalje, cilj je i odrediti individualne razlike među sudionicima, utvrditi univerzalnosti i posebnosti u odnosu na jezični razvoj u drugim jezicima (istraživanja međujezičnih usporedbi) te raščlaniti utjecaj ulaznoga jezika, odnosno jezika odraslih na razvoj djetetova jezika. Takve spoznaje nisu izravno dostupne, već uključuju posebne vrste analize strukturirane metodologijom znanstvenih istraživanja.

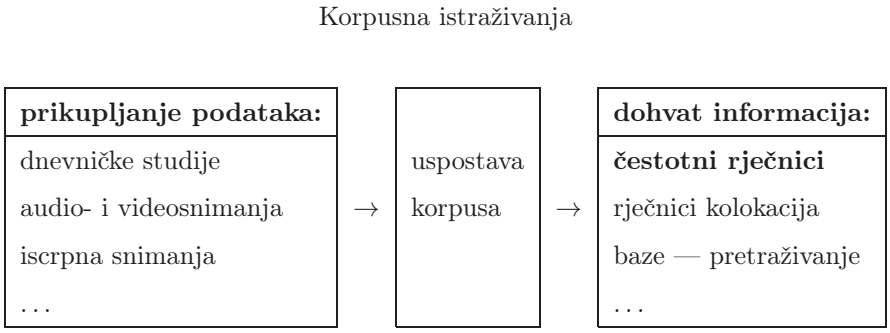
HKDJ za sada sadrži 126 prijepisa zvučnih zapisa troje djece. Snimljeni su i govorni uzorci blizanaca i djece dvojezičnih govornika, ali ti su zvučni zapisi još u postupku stvaranja prijepisa. HKDJ sadrži ukupno 87 396 iskaza od čega su 40% dječji, a preostalih 60% iskazi odraslih koji su upućeni djeci (Kuvač i Palmović 2007: str. 149). Prijepisi su oblikovani prema pravilima CHAT-a te obrađeni u računalnom programu CLAN. Nakon primarne obrade podatci su obrađeni na morfološkoj razini (više o obradi hrvatskoga korpusa vidi u Kuvač, Palmović 2001, 2007). Tek na morfološki oblikovanim prijepisima mogu se dohvatiti valjani podatci o jezičnome razvoju.

1.2. Hrvatski čestotni rječnik dječjega jezika (DjeČeR)

Hrvatski čestotni rječnik oblikovan je na temelju znanostvenoistraživačkih spoznaja kako bi omogućio rasvjetljavanje temeljnih pitanja leksičkoga razvoja u usvajanju hrvatskoga jezika. Postavljena su pitanja na koja bi rječnik svojom strukturom trebao odgovoriti: a) koji je temeljni rječnik djece do treće godine, b) kakav je morfološki opis temeljnoga rječnika i c) koje su osnovne morfosintaktičke kategorije usvojene u jeziku djeteta do tri godine.

HKDJ je, kao specifičan korpus govornoga jezika, prikupljan kako bi se omogućile spoznaje o ranomu jezičnomu razvoju. Korpus je javno dostupan, međutim, sama dostupnost takva mnoštva podataka nije uvijek dovoljna za njegovu raširenu primjenu. Da bi se analizirao korpus, potrebno je znati se služiti specijaliziranim računalnim programima na relativno naprednoj

razini. Čak i uz to, potreban je određeni rok i nekoliko stupnjeva analize da bi se došlo do određenih podataka. Primjenom inovativnih znanstvenometodoloških rješenja stvoren je izvor koji omogućuje krajnjim korisnicima lakše i brže pronalaženje potrebnih podataka u korpusu. Pri tome se krajnjim korisnicima smatraju znanstvenici i stručnjaci iz jezične prakse (od primijenjenih lingvista do logopeda), koji za dohvat podataka više ne moraju ovladavati primarno znanstvenometodološkim sredstvima analize korpusa. Dohvat isključuje takvu analizu te zahvaljujući svojoj strukturi brzo i precizno odgovara na postavljena pitanja.



Slika 1. Postupak dohvata podataka u korpusnim istraživanjima

Čestotni rječnici govornoga jezika nisu česti i postoje za ograničen broj jezika, npr. engleski — moguće je naručiti ispisne čestotne liste podkorpusa govornoga jezika najvećega korpusa američkoga engleskoga (*Corpus of Contemporary American English* — Davies 2008), njemački (Jones i Tschirner 2006). Još su rjeđi čestotni rječnici dječjega jezika, naročito samostalni. Primjerice, za švedski jezik postoji korpus govornoga jezika koji sadrži i potkorpus dječjega jezika te je na temelju njega izrađen čestotni rječnik (Allwood 1996. i novija izdanja). No, čestotni podatci nisu odvojeni prema potkorpusima te tako ovaj izvor ne predstavlja čestotni rječnik dječjega jezika, iako se podatci mogu neizravno iščitati (uz svaku natuknicu dostupna je i pojavnost u određenom potkorpusu). Samostalni čestotni rječnici dječjega jezika izrazito su rijetki. Dijelom je razlog tome zahtjevnost njihove izrade, posebice za morfološki bogate jezike, a dijelom omogućavanje dostupnosti podataka o čestotnosti drugačijim alatima (primjerice, različitim pretraživačima baza podataka, primjerice CLEX — *Cross Linguistic Lexical Norms* — dostupno na http://www.cdi-clex.org/, Nørgaard Jørgensen i sur. 2009).

2. Postupak izrade DjeČeRa

DjeČeR je razvijen u skladu s osobinama hrvatskoga jezika s korpusom govornoga dječjega jezika kao temeljnim izvorom informacija. Temeljni postupci strukturiranja podataka i izrade DjeČeRa razvijeni su kako bi odgovorili na znanstvenoistražvačka pitanja o dječjem rječničkom razvoju te kako bi pružili podatke ključne za praktični rad. Mogućnost izrade različitih metodoloških sredstava koja omogućuju lakši dohvat podataka dio je procesa korpusnih istraživanja. Kako bi se na toj osnovi mogla izraditi različita metodološka sredstva, preduvjeti za to moraju biti ugrađeni već u samu izradu korpusa. Upravo su zato prvi preduvjeti za nastanak DjeČeRa morali biti planirani i zadovoljeni još pri izradi Hrvatskoga korpusa dječjega jezika.

2.1. Morfološko kodiranje — preduvjet analize čestotnosti

Neposredno nakon transkripcije i osnovnoga kodiranja datoteka HKDJ-a pristupilo se morfološkoj analizi. Ona je nužna iz više razloga. Jedan je povezivanje oblika koji se pojavljuju u prijepisu s osnovnim oblicima riječi pogodnima za analizu i pretraživanje prema čestotnosti. Drugi je izbjegavanje efekta istozvučnosti u određivanju čestotnosti.

Konkretno, u osnovnom se prijepisu riječi pojavljuju u različitim fleksivnim oblicima. Računalo pri tome pri prebrojavanju kao riječ definira prostor između dviju praznina. Ako su nizovi znakova unutar dvije praznine identični, računalo će ta dva niza brojati kao istu riječ (različnicu) koja se ponavlja više puta (ima više pojavnica). Na primjer, u primjeru u (1) dva se puta ponavlja niz znakova *'snimamo'*. Dakle, različnica *'snimamo'* pojavila se u dvije pojavnice. Međutim, ona ni na koji način nije povezana s različnicama *'snimi'* ili *'snimila'*. Morfološka analiza omogućava da se ti različiti oblici vezuju uz zajednički osnovni oblik riječi (*'snimiti'*), to jest na natuknicu umnoga rječnika. Osim toga, kako bi se izbjegla gramatička i leksička istozvučnost, različnice se uslijed morfološke analize opisuju tako da se istoznačnost izbjegne. Tako opisani oblici nazivaju se obličnice (Jelaska, 2005). Na primjer, u primjeru u (1) četiri se puta ponavlja različnica *'snimi'*. U tekstu je svaki puta riječ o drugomu licu imperativa. Međutim, taj isti niz znakova može označavati i 3. lice jednine prezenta. Morfološka analiza zato će toj jednoj različnici pripisati dvije različite obličnice i povezati ih s pripadajućom natuknicom.

(1) Sadržajni retci osnovne datoteke

*MAR: zašto si tu donesla@d?
*SAN: da malo snimamo šta se razgovaramo.
*MAR: daj me snimi.

*SAN: evo sad snimamo # (h)oćeš slušat?
*MAR: aha.
*SAN: xxx.
*MAR: sad snimi [//] sad snimi [//] sad to snimi.
*SAN: bravo tak(*2) # pusti(*4) sad ovdje # evo ti na.
*MAR: ja sam sad tebe snimila.

Kako izgleda morfološki opis jednoga dijela datoteke može se vidjeti u primjeru u (2).

(2) Dio morfološki kodirane datoteke

*MAR: zašto si tu donesla@d ?
%mor: ADV|zašto
V:AUX:IMPF|biti&PRES:2S:CLIT
ADV|tu V:1:PFV:TRANS|donijeti&PART-FEM:SG ?
*SAN: da malo snimamo šta se razgovaramo .
%mor: CONJ:SUBOR|da
ADV|malo
V:5:IMPF:TRANS|snimati&PRES:1P
PRO:REL|što&NOM
PRO:REFL|sebe&ACC:SG:CLIT V:5:IMPF:REFL|
|razgovarati&PRES:1P .
*MAR: daj me snimi .
%mor: V:5:PFV:TRANS|dati&IMP:2S
PRO:PERS:1S|ja&ACC:CLIT
V:4:PFV:TRANS|snimiti&IMP:2S .
*SAN: evo sad snimamo # (h)oćeš slušat ?
%mor: CO|evo ADV|sad V:5:IMPF:TRANS|snimati&PRES:1P
V:1:IMPF:TRANS:MOD|htjeti&PRES:2S V:5:IMPF:TRANS|
|slušati&INF ?

2.2. Morfološko kodiranje u programu CLAN

Programski paket CLAN sadrži nekoliko programa od kojih je za morfološku analizu, to jest za analizu morfološkoga razvoja snimane djece, ključan program za morfološku obradu prijepisa MOR. Cilj je te obrade stvarati posebne vrste datoteka, *.mor* datoteke, u kojima se svaka izgovorena riječ morfološki opisuje u zasebnoj liniji (%mor), kao što je vidljivo u Primjeru 2. Takve se datoteke poslije mogu obrađivati ostalim CLAN-ovim programima i dobiti podatke, na primjer, koliko je puta upotrebljen određeni padež ili u kojem se obliku najčešće pojavljuje određeni glagol.

Međutim, originalni CLAN omogućava takvu morfološku obradu samo prijepisa na engleskom jeziku. Istraživači ostalih jezika moraju prilagoditi

CLAN svom jeziku kako bi rabili tu opciju (MacWhinney 2008). Prilagodbu je moguće načiniti na dva načina: zadavanjem niza pravila prema kojima program određuje morfološke osobine određene obličnice (npr. u engleskome jedno od takvih pravila može biti izolacija morfema *-ed,* kao oznake prošloga vremena) ili ugrađivanjem niza već morfološki opisanih obličnica — program taj popis uspoređuje s prijepisom i pronađene pojavnice prepoznaje i pridružuje im opis.

Razlikovanje pojavnica i različnica automatizirano je u CLAN-u. Međutim, prepoznavanje pripadnosti natuknici složenije je i ovisi o navedenim načinima prilagodbe. Ako je programu dan niz pravila, taj niz pravila mora moći i odrediti kako povezati različnicu s natuknicom (npr. u engleskom oduzimanjem morfema *-ed*). Naravno, u slučaju iznimaka (na primjer, pojavnica *weed*) one se mogu zasebno definirati. Ako je programu dostavljen niz opisanih različnica, te različnice mogu u svom opisu sadržavati natuknicu kojoj pripadaju.

Kako u hrvatskom postoji velik broj homonimnih morfema (npr. *-a* označava imenicu ženskoga roda u nominativu jednine, 3. lice prezenta, nominativ množine srednjega roda imenica, genitiv muškoga roda imenica /živo/, ženski rod pridjeva i dr.), složena pravila alternacija fonema i mnogo glagola netransparentne morfologije (npr. netematska glagolska vrsta), prilagodba CLAN-a hrvatskomu jeziku provedena je na drugi način. Jezici koji su po takvim ili sličnim morfološkim osobinama srodni hrvatskom (npr. litavski) također su prilagodbu proveli na takav način. Iznimka je u tome ruski jezik kod kojega su se istraživači dječjega jezika odlučili za pravila zato jer je već postojao definiran niz pravila za ruski jezik. Za samu prilagodbu tih postojećih pravila CLAN-u bile su potrebne oko tri godine (Voekova 2008, osobni razgovor).

Iako je nakon prilagodbe programa morfološka analiza značajno olakšana, još je potrebna stručna pomoć istraživača u određivanju istozvučnh pojavnica. Prolaskom datotekom istraživaču se nude mogućnosti morfološkoga opisa među kojima mora izabrati ispravan. Tek nakon toga morfološke su datoteke spremne za dalju analizu. Prilagodbu CLAN-a osmislio je i proveo Marijan Palmović, a morfološku su obradu proveli Marijan Palmović i Gordana Hržica. Obrada je rezultirala nizom *.mor* datoteka, po jednom za svaki mjesec snimanja (Hržica 2010).

2.3. Pripremanje podataka za Hrvatski čestotni rječnik dječjega jezika

Kako bi se dobili podatci pogodni za dalju obradu, morfološke su datoteke svakoga djeteta pretvorene u tabličnu bazu. Svaki redak tablice odgovara obličnici koju je određeno dijete izreklo u određenom mjesecu svoje krono-

loške dobi. Svaka je obličnica vezana uz određenu natuknicu, a uz obličnicu su dodani podatci o čestotnosti i morfosintaktički opis. Dodatan je posao bio potreban kako bi se ujednačile neke razlike u morfološkom kodiranju do kojih je došlo zbog dugotrajnosti projekta. Naime, tijekom godina razvoja HKDJ-a na njemu je radilo više ljudi koji su u obradu podataka unijeli i neke individualne razlike, uključujući, na primjer, različit redoslijed upisivanja kratica gramatičkih kategorija, različite kratice, uporabu različitih gramatičkih opisa kao osnovu kodiranja i slično. Ujednačenost takvih razlika nužan je preduvjet dalje obrade podataka.

2.4. Analiza tablične baze provedbom upita

Kako bi se iz tablične baze mogli dobiti potrebni podatci, razvijen je niz upita. Upiti su posebni naputci koji omogućuju različite načine svrstavanja i organizacije podataka dostupnih u tabličnoj bazi. Na primjer, jedan od upita mogao bi biti *prikaži natuknice sve troje djece razvrstane abecedno*.

Upiti su sastavljeni tako da podatci dostupni u bazi mogu što lakše odgovoriti na osnovna pitanja koja bi mogao postaviti istraživač dječjega jezika ili stručna osoba kojoj su u radu nužni podatci o dječjem jeziku. Provedbom različitih upita formirani su različiti dijelovi Hrvatskoga čestotnoga rječnika dječjega jezika, to jest, određena je njegova struktura.

3. Struktura hrvatskog čestotnog rječnika dječjega jezika

Struktura DjeČeRa polazi od temeljnih pitanja koja se postavljaju u literaturi o ranomu jezičnomu razvoju. Osim znanstvenicima, odgovori su na ta pitanja izravno potrebna i velikom broju stručnjaka iz stručne djelatnosti koja se veže uz rad na dječjem jeziku. Neizravno su odgovori bitni i kako bi se poznavao tijek jezičnoga razvoja, normirali jezični testovi i uspostavili orijentiri u stručnome radu s dječjim jezikom.

3.1. Ciljevi DjeČeRa

Primjenom posebnoga načina organizacije podataka dostupnih u korpusu, to jest organizacijom podataka u čestotni rječnik, pokušalo se omogućiti brže, izravnije i temeljitije rasvjetljavanje triju temeljnih pitanja leksičkoga razvoja: opis razvoja temeljnoga rječnika, morfološki opis temeljnoga rječnika i opis usvajanja osnovnih morfosintaktičkih kategorija u dječjem jeziku.

3.1.1. Rječnički razvoj — opis razvoja temeljnoga rječnika

Pod pojmom temeljni rječnik smatra se popis natuknica koje djeca usvoje do treće godine života. Uspostavom temeljnoga rječnika barem nekoliko djece

dobila bi se okvirna slika rječnika u ranomu razdoblju početnoga razdoblja usvajanja hrvatskoga jezika. Takav bi rječnik služio kao dobra osnova za dalji znanstveni (npr. semantička analiza riječi i kategorija, struktura rječnika) i stručni rad (npr. pomagalo u odabiru riječi za ispitne materijale, materijale za jezičnu terapiju i slično). Osim utvrđivanja temeljnoga rječnika, istraživači dječjega jezika, ali i stručnjaci koji se njihovim saznanjima koriste, htjeli bi znati kako se taj rječnik mijenja, odnosno kako raste s povećanjem kronološke dobi djeteta. Pojedinčeve razlike u jezičnomu (pa tako i leksičkomu) razvoju mogu biti izrazite te je stoga potrebno uzeti u obzir njihov utjecaj na spoznaje o temeljnomu rječniku.

3.1.2. Morfološki opis temeljnoga rječnika

Pridruživanje kategorija morfološkoga opisa natuknicama temeljnoga rječnika pruža dodatne obavijesti o njegovu ustroju i strukturi s obzirom na pripadnost različitim gramatičkim kategorijama. Gramatičke kategorije temeljnoga rječnika također su ovisne o kronološkoj dobi. Praćenje promjena u sastavu rječnika s obzirom na pripadnost gramatičkim kategorijama važno je kako bi se stekao uvid u razvojne promjene u strukturi rječnika. Uz praćenje promjena vezanih uz kronološku dob potrebno je u obzir uzeti moguće individualne razlike među pojedinim govornicima.

3.1.3. Opis usvajanja osnovnih morfosintaktičkih kategorija

Temeljni rječnik, točnije, oblici (obličnice) koje se pojavljuju pod određenim natuknicama temeljnoga rječnika, pružaju i posredne podatke o usvojenim morfosintaktičkim kategorijama djece do treće godine. Naime, ako se uz svaku natuknicu ispišu i njezine obličnice, može se pratiti koje se gramatičke kategorije pojavljuju uz određenu natuknicu i šire, na primjer, uz natuknice koje pripadaju u određenu vrstu riječi.

Pretpostavlja se da će broj različitih morfosintaktičkih kategorija rasti s obzirom na kronološku dob, to jest s obzirom na pretpostavljeni razvoj jezika u skladu s kronološkom dobi. Pojedinčev jezični razvoj može dovesti i do razlika u usvojenosti određenih gramatičkih kategorija kod različite djece, u praćenju razvojne komponente, ali i u konačnomu rezultatu.

3.2. Struktura DjeČeRa

Navedeni znanstvenoistraživački ciljevi u istraživanjima usvajanja dječjega jezika utjecali su na oblikovanje strukture DjeČeRa. Kako bi se olakšalo pretraživanje čestotnosti natuknica i oblika u HKDJ-u u skladu s uobičajenim ciljevima proučavanja jezičnoga usvajanja, rječnik je podijeljen na dva osnovna dijela.

3.2.1. Prvi dio rječnika — popisi natuknica prema različitim kriterijima

U prvom dijelu navedene su natuknice raspoređene prema trima kriterijima: općenito niz natuknica, natuknice grupirane s obzirom na kronološku dob te natuknice grupirane s obzirom na vrstu riječi kojoj pripadaju, kao što je prikazano na slici 2.

1. dio rječnika — natuknice

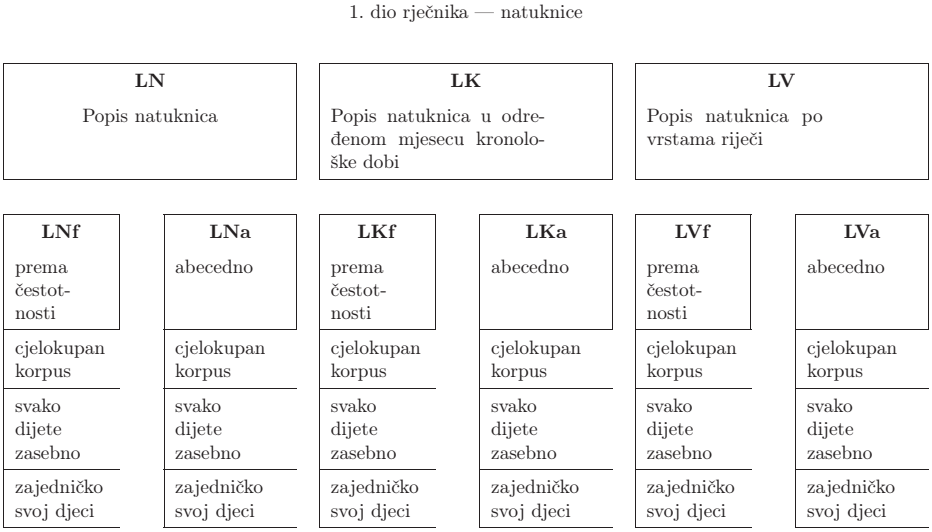
LN			**LK**				**LV**	
Popis natuknica			Popis natuknica u određenom mjesecu kronološke dobi				Popis natuknica po vrstama riječi	
Primjer:			Primjer:				Primjer:	
5 najčešćih natuknica			4 natuknice u kd 1;5				4 najčešća prijedloga	
3653	biti	glagol	*1;5*				*prijedlog*	
1895	a	veznik	9	batić	imenica	12	397	u
1813	ja	zamjenica	7	ja	zamjenica	689	222	na
1730	ovaj, -a, -o	zamjenica	6	dati	glagol	192	128	sa
1630	htjeti	glagol	4	sam, -a, -o	zamjenica	14	85	za

Slika 2. Osnovna trodijelna struktura prvoga dijela DjeČeRa

Natuknice su u svakoj od navedene tri kategorije svrstane s obzirom na dva osnovna kriterija: čestotnost i abecedni poredak, kao što je prikazano na slici 3. To znači da se unutar svakoga od tri osnovna dijela mogu pratiti natuknice prema ta dva kriterija. Na primjer, unutar treće kategorije (LV) možemo pratiti kojih je pet najčešćih glagola ili možemo određeni glagol pronaći u abecednomu popisu i utvrditi njegovu čestotnost.

Slika 3. Svrstavanje unutar osnovne trodijelne strukture prvoga dijela DjeČeRa

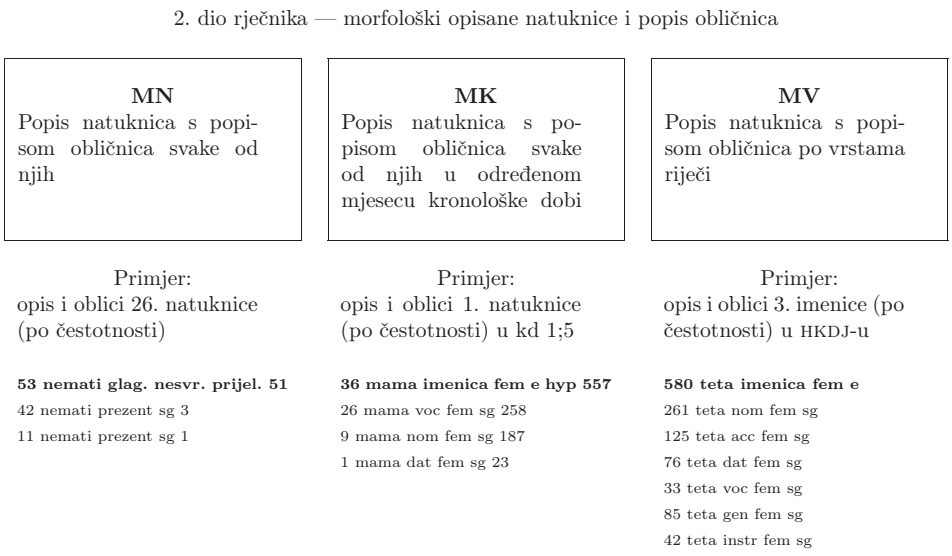
Unutar ova tri dijela, kako bi se sačuvali i podatci o pojedinačnim razlikama, natuknice su dodatno prikazane na tri načina (slika 4): ukupno unutar određene kategorije (dakle, sve natuknice koje su se pojavile kod bilo kojega djeteta), pojedinačno za svako dijete te natuknice koje su se pojavile u korpusu svakoga djeteta. Prednosti takva načina organizacije podataka mogućnost je izravnijega pretraživanja. Na primjer, unutar druge kategorije (LK) može se saznati koje su se natuknice u određenomu mjesecu pojavile ukupno u HDKJ-u, koje su se u tom istom mjesecu pojavile kod pojedinoga djeteta te koje su se u tom mjesecu pojavile kod svakoga od troje djece.



Slika 4. Detaljna struktura prvoga dijela DjeČeRa

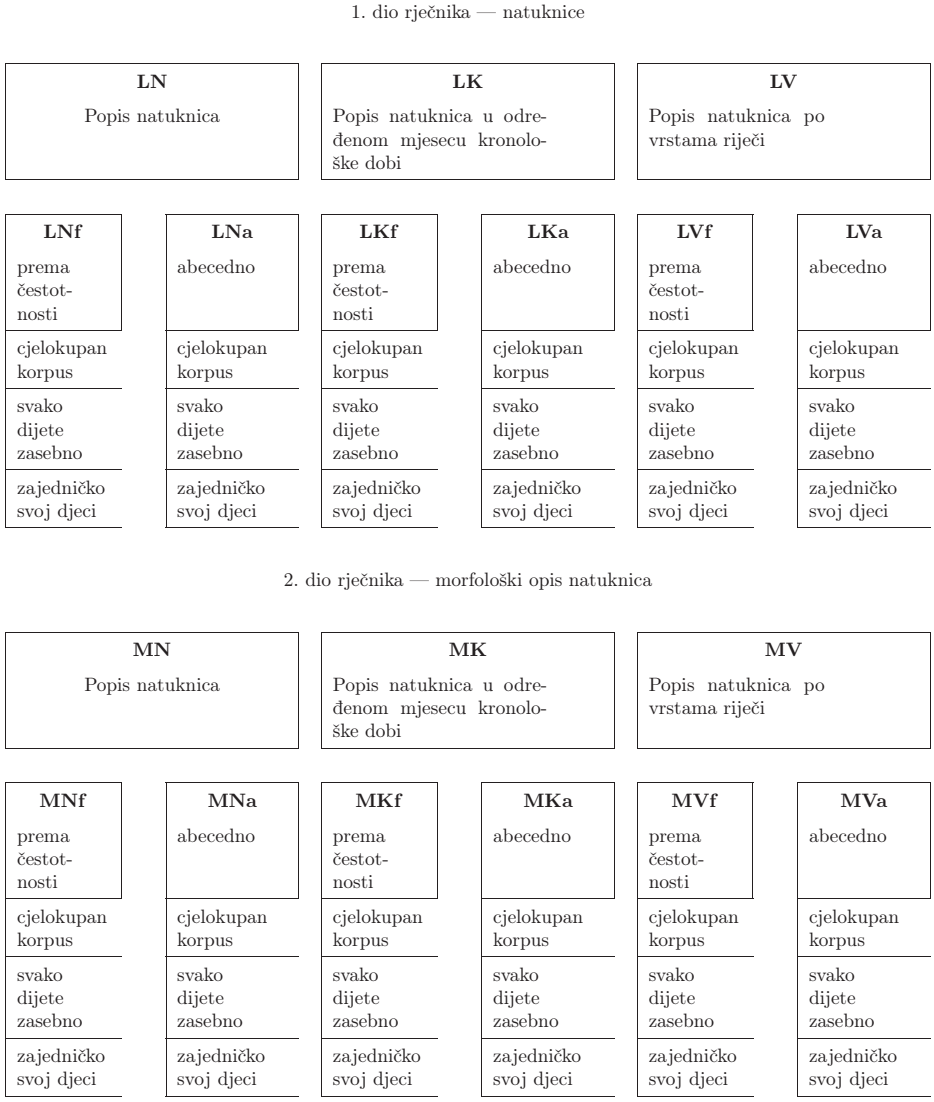
3.2.2. Drugi dio rječnika — popisi morfološki opisanih natuknica prema različitim kriterijima

U drugomu dijelu rječnika ponavljaju se natuknice već prikazane u prvom dijelu. Načini njihova svrstavanja ostaju isti (slika 5). Natuknicama se dodaje morfološki opis. Morfološki se opis sastoji od dvije razine. Svakoj se natuknici pridodaju precizniji opisi gramatičkih kategorija (npr. glagolima se pridodaje pripadnost glagolskoj vrsti i razredu, vid, prijelaznost). Uz svaku je natuknicu i točan popis (i čestotnost) oblika koji su se pojavili u HKDJ-u. Tako se npr. može vidjeti u kojim se točno oblicima pojavio glagol *'željeti'*. Razlog je ovakva dvostrukoga navođenja natuknica (u prvom dijelu bez morfološkoga opisa, u drugom s opisom) lakše snalaženje u rječniku. Naime, morfosintaktičke informacije značajno otežavaju preglednost popisa pojmova, a nisu nužne svim korisnicima rječnika. Autori su stoga odlučili omogućiti zasebno pretraživanje tih dvaju vrsta podataka.



Slika 5. Osnovna trodijelna struktura drugoga dijela DjeČeRa

Kao i prvi dio rječnika, i ovaj dio zadržava načine svrstavanja prema abecedi i prema čestotnosti. Prikazani su podatci koji se pojavljuju u cjelokupnom korpusu, u korpusu svakoga pojedinoga djeteta te podatci zajednički svoj djeci (slika 6). Takav usporedni prikaz omogućava iscrpan i brz pregled podataka bez potrebe za izravnom analizom prijepisa u HKDJ-u.



Slika 6. Detaljna struktura DjeČeRa

4. Struktura DjeČeRa kao odgovor na istraživačka pitanja o rječničkom razvoju u usvajanju hrvatskoga jezika

Hrvatski čestotni rječnik dječjega jezika omogućava brz i jednostavan dohvat podataka iz Hrvatskoga korpusa dječjega jezika koji su bitni kao odgovori na pitanja o leksičkom razvoju postavljenima kao cilj razvoja ovoga izvora.

4.1. Rječnički razvoj — razvoj temeljnoga rječnika

DjeČeR u svojim različitim dijelovima daje dobar pregled strukture temeljnoga rječnika troje djece, uz mogućnost uvida u razvojnu sastavnicu (prikazom po mjesecima kronološke dobi). Omogućena je i usporedba pojedinačnih razlika s obzirom na opći korpus i na korpus zajednički za svo troje djece.

4.2. Morfološki opis temeljnoga rječnika

U morfološkomu dijelu rječnik omogućuje uvid u dublju razinu analize morfološkim opisom natuknica. Uz to se navođenjem obličnica koje se pojavljuju unutar pojedine natuknice omogućava i nova razina analize. Naime, ne samo da je vidljivo u kojim se oblicima pojavljuju određene natuknice, nego to navođenje doprinosi uvidu u usvojenost određenih gramatičkih kategorija. Uz prisutnost razvojne komponente prikaza (po mjesecima kronološke dobi) moguće je posredno pratiti pojavljivanje gramatičkih kategorija unutar određene natuknice, određene vrste riječi ili detaljnijih kategorija (npr. imenica srednjega roda). Uz mogućnost uvida u pojedinačne razlike, DjeČeR omogućava donošenje nekih općenitih zaključaka ne samo o leksičkom, već i o gramatičkom razvoju djece do treće godine.

4.3. Usvajanje osnovnih morfosintaktičkih kategorija

Iako DjeČeR omogućava posredni uvid u usvajanje osnovnih morfosintaktičkih kategorija (analizom određenih natuknica), ovakav način organizacija podataka ne može pružiti brze odgovore na pitanja kao što su koji se padeži pojavljuju u imenicama ženskoga roda do druge godine ili koja se glagolska vremena pojavljuju do treće godine. Nastavak rada na Hrvatskomu korpusu dječjega jezika pokušat će ponuditi upravo odgovore na takva pitanja. Međutim, tiskani medij koji bi na toliko načina svrstava podatke ne bi bio racionalan. Upravo je zato nastavak DjeČeRa zamišljen kao javno dostupna baza s mogućnošću pretraživanja po temeljnim morfološkim kategorijama.

4. Zaključak

Hrvatski čestotni rječnik dječjega jezika strukturiran je temeljem spoznaja o potrebama istraživanja dječjega jezika uporabom metodologije korpusnih istraživanja. Temelji se na izvorima i alatima već razvijenima za hrvatski jezik (Hrvatski korpus dječjega jezika te njegov morfološki leksikon). Struktura mu je određena primarno znanstvenim spoznajama o ključnim aspektima razvoja rječnika u hrvatskom jeziku.

Temeljna je svrha DjeČeRa strukturirati već dostupne podatke tako da omogući brz i lak dohvat podataka, služeći kao izvor podataka pri razvoju znanstvenoistraživačkih materijala (primjerice, razvoju istraživačkih paradigmi), razvoju ispitnoga materijala (prilagodbi i proizvodnji jezičnih i drugih testova) te u svakodnevnom stručnom radu (primjerice, pri razvoju materijala za rad tijekom jezične terapije, nastave hrvatskoga jezika i slično). Drugim riječima, DjeČeR ima svoju ulogu u svakoj prigodi u kojoj su potrebni nepristrani i pouzdani podatci o prisutnosti i čestotnosti određene natuknice i njezinih oblična u dječjemu jeziku.

5. Literatura

Allwood, J., ur. (1996. i novija izdanja) Talspråksfrekvenser, Ny och utvidgad upplaga, *Gothenburg Papers in Theoretical Linguistics S21,* Göteborg: Göteborg University, Department of Linguistics.

Behrens, H. (2008) Corpora in language acquisition research: History, methods, perspectives, u H. Behrens (ur) *Corpora in Language Acquisition Research: History, methods, perspectives,* Amsterdam: John Benjamins Publishing Company.

Bujas, Ž. (1975.a) *Ivan Gundulić „Osman“* - Komputorska konkordancija, Zagreb: Sveučilišna naklada Liber.

Bujas, Ž. 1(975.b) Computers in the Yugoslav Serbo-Croatia : English Contrastive Project, *Bilten Instituta za lingvistiku,* 1, 44–58.

Crystal, D. (2003) *English as a Global Language,* Cambridge: Cambridge University Press.

Davies, M. (2008) *The Corpus of Contemporary American English: 450 million words, 1990-present,*

URL: http://corpus.byu.edu/coca/

Kucera, H, Nelson., F.W. (1967) *Computational analysis of present-day American English,* Providence: Brown University Press.

Furlan, I. (1961) *Raznolikost rječnika. Struktura govora,* Zagreb: Filozofski fakultet Sveučilišta u Zagrebu (doktorska disertacija).

Hržica, G. (2010) P*ojavljivanje morfoloških kategorija i sintaktičkih obrazaca glagola u usvajanju hrvatskog jezika.* Zagreb: Filozofski fakultet Sveučilišta u Zagrebu (kvalifikacijski rad).

Jelaska, . i sur. (2005) *Hrvatski kao drugi i strani jezik,* Zagreb: Hrvatska sveučilišna naklada.

Jones, R., Tschirner, E. (2006) *A frequency dictionary of German,* London: Routledge.

Kovačević, M.(2002) *Hrvatski korpus dječjega jezika,* CHILDES project.

Kuvač, J., Palmović, M. (2001) Računalna obrada dječjega jezika na primjeru usvajanja umanjenica, *Suvremena lingvistika*, 50/51, 101–111.

Kuvač, J., Palmović, M. (2007) *Metodologija istraživanja dječjega jezika,* Jastrebarsko: Naklada Slap.

MacWhinney, B. (2000) *The* CHILDES *project: Tools for analyzing talk. Third Edition,* Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2008) Enriching CHILDES for morphosyntactic analysis, u H. Behrens (ur.) *Corpora in Language Acquisition Research: History, methods, perspectives,* Amsterdam: John Benjamins Publishing Company.

MacWhinney, B., Snow, C. (1985) The child language data exchange system, *Journal of Child Language* 1; 271–295 .

Nørgaard Jørgensen, R., Dale, P. S., Bleses, D., Fenson, L. (2009) CLEX: A cross-linguistic lexical norms database, *Journal of Child language,* 37, 2; 419–428.

Svartvik, Jan, ur. (1992) *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm.* Berlin: Mouton.

Tadić, M. (2003) *Jezične tehnologije i hrvatski jezik,* Zagreb: Exlibris.

Croatian Frequency Dictionary of Child Language

*Nowadays language corpora are recognised as valuable and informative sources of linguistic information. However, retrieving the available data can be demanding and complex, therefore sometime not suitable for all users that could benefit from it. The only existing Croatian corpus of spoken language is the Croatian Corpus of Child Language (*CCCL *— Kovacevic, 2002). Speech samples were taken from three children, in equable time periods, from the onset of speech to three years of age. Samples were transcribed according the rules of* CHAT*, using the computer programme* CLAN*.* CCCL *is available on-line in the* CHILDES *(Child Language Data Exchange System — http://childes.psy.cmu.edu/data/Slavic/). It is designed to provide data about lexical and grammatical development in language acquisition. Consequently, a Croatian frequency dictionary of child language (*CFDCL*) has been designed to enable easier data retrieval form* CCCL*. It allows the analyses of most frequent lemmas in all three sub-corpora according to frequency, alphabetic ordering, time of appearance, and part-of-speech. Furthermore, it preserves the morphological encoding of types, and number of types and tokens. Therefore it incorporates a larger amount of information than traditional corpora of written language, enabling users to extract relevant information about child language development such as type/token ratio, lexical diversity, morphological diversity, etc.*

Key words: Croatian corpus of child language, Croatian Frequency Dictionary of Child Language, CHILDES, lemmatization, tagging of corpus, structure of CFDCLP

Ključne riječi: Hrvatski korpus dječjega jezika, Hrvatski čestotni rječnik dječjega jezika, CHILDES, lematizacija, označavanje korpusa, struktura čestotnoga rječnika